DOCUMENT RESUME

ED 336 407 TM 017 122

AUTHOR	Sinclair, Norma; Pecheone, Raymond L.
TITLE	Empirical Appraisal of Items for Innovative Teacher Tests.
PUB DATE	Apr 91
NOTE	17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
AVAILABLE FROM	Norma Sinclair, Bureau of Research and Teacher Assessment, Connecticut State Department of Education, P.O. Box 2219, Hartford, CT 06145.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Beginning Teachers; Elementary Education; *Elementary School Teachers; Higher Education; *Item Response Theory; *Licensing Examinations (Professions); Mathematical Models; State Programs; *Teacher Certification; Teacher Education; Teaching Methods; Test Construction; Testing Programs; *Test Items
IDENTIFIERS	*Connecticut Elementary Certification Test; One Parameter Model; Subject Content Knowledge; Two Parameter Model; Unidimensionality (Tests)

ABSTRACT

The impact of test item multidimensionality was examined as it affected fitting items on a teacher licensure test to an item response theory (IRT) model. Test item data from the 1990 study of G. W. Guiton and G. Delandshere are used. The Connecticut Elementary Certification Test (CONNECT) is a licensure examination designed to assess the subject matter knowledge of a beginning elementary teacher in the state. Nine content areas and pedagogy are covered. Items selected for analysis were taken from two test forms of the CONNECT administered prior to the date when it became a requirement for certification. In all, 826 responses were received from 681 candidates (some took both test forms). Classical item statistics indicated that test items were relatively heterogeneous, with the highest frequency of point biserials in the 0.10 to 0.30 range. These statistics confirm the previous (1990) finding of more than one underlying construct. However, successful attempts to fit the data to one-parameter and two-parameter models support the contention that a dominant underlying factor is sufficient to meet the assumption of unidimensionality required for item response analyses. An appendix summarizes the relationship between content knowledge and pedagogical knowledge for teachers. Three tables present study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NORMA SINCLAIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED336407

EMPIRICAL APPRAISAL

OF ITEMS FOR INNOVATIVE TEACHER TESTS

Norma Sinclair
Connecticut State Department of Education

Raymond L. Pecheone
Connecticut State Department of Education

A paper prepared for the 1991 Annual Meeting of the National Council of Measurement in Education in Chicago. Requests for copies should be sent to Norma Sinclair, Bureau of Research and Teacher Assessment, Connecticut State Department of Education, P.O. Box 2219, Hartford, CT 06145.

TM017122

BEST COPY AVAILABLE

Teacher licensure tests are undergoing a fundamental metamorphosis. In response to criticisms of the triviality of content, test developers have directed their efforts at building items that are more sensitive to the applied skills teachers need as a result of classroom experience. These efforts have resulted in more contextualized content and pedagogy items designed to reflect the richness and complexity of actual teaching practices.

The development of elementary education licensure tests more sensitive to teaching practices pose a particular problem. With the broad scope of subject matter elementary teachers are required to teach, licensure examinations for elementary teachers currently, must assuredly assess knowledge across multiple content areas. This quality of elementary education licensure tests have resulted in examinations that are multidimensional in content.

Given the multidimensionality of elementary education licensure examinations, test developers frequently assume that the application of item response theory is out of the question. By contrast, however, dimensionality is both a content and empirical issue. Accordingly, rather than make assumptions regarding the dimensionality of test items, Guiton and Delandshere (1990) examined empirical evidence derived from pilot test data of an elementary education test, the Connecticut Elementary Certification Test (CONNECT). Their investigations resulted in the identification of a dominant general factor. A secondary factor, supporting the existence of content-specific pedagogical knowledge, was also identified.

Admittedly, the Guiton and Delandshere findings failed to identify the single underlying construct preferred to meet the unidimensionality assumption of item response. However, Humphreys (1984) and Hambleton and Swaminathan (1985) suggest that a single dominant dimension or factor may be sufficient to meet the assumption. These findings informed the decision to empirically assess test items for the CONNECT, using classical and item response methods.

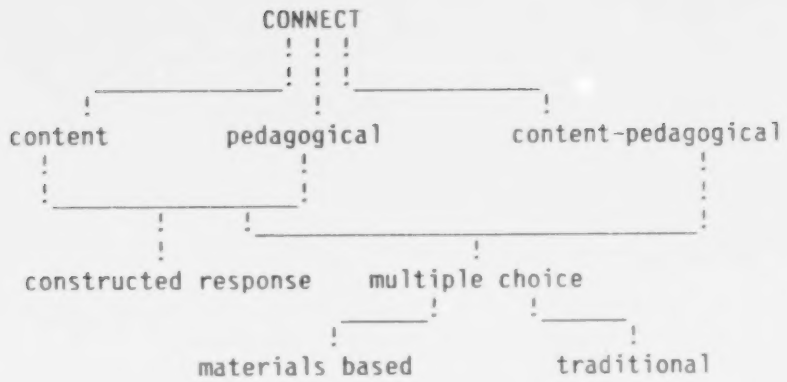
The purpose of this paper is to examine the impact of test item multidimensionality as it affects fitting test items to an IRT model. In addition, the paper will examine issues arising from efforts to fit such data to a model. In effect, this paper represents an extension of the Guiton and Delandshere (1990) investigation. Test item data developed for the same licensure test addressed by the 1990 paper will be used.

Description of Test Instrument

The Connecticut Elementary Certification Test (CONNECT) is a licensure examination designed to assess the subject matter knowledge expected of a beginning elementary education teacher in the state of Connecticut. A passing score on the CONNECT is one of the requirements for the issuance of an initial Connecticut elementary teaching certificate.

In an effort to respond to validity considerations described previously, items were developed along three dimensions: the format of the item, the type of knowledge to be assessed and the content-area context for the item (figure 1).

FIGURE 1


CONNECT
content
pedagogical
content-pedagogical
constructed response
multiple choice
materials based
traditional

Item Format

Each test form comprises a 100-item multiple-choice component and a 10-item constructed response component developed from 61 objectives. Approximately 10 percent of the multiple-choice items are materials based; that is they require the review and analysis of materials (work sheets, IEPs, student performance records) a beginning teacher is likely to encounter frequently. All of the constructed response items are based on videotape vignettes of activities likely to take place in an elementary school classroom. Only the multiple-choice items will be addressed in this paper.

Types of Knowledge

In developing item specifications and the items themselves, Connecticut attempted to measure the types of teacher knowledge defined by Shulman (1987)--content knowledge, pedagogical knowledge and content-specific pedagogical knowledge. (Definitions may be found in Appendix A.) Items developed to measure content-specific pedagogical knowledge primarily represent the endeavor to include contextualized items that reflect actual practice.

Content Areas

The CONNECT assesses subject matter knowledge in nine content areas as well as pedagogy. Those content areas comprising the core of the examination consist of language arts, mathematics, science and social studies. Additional content areas consist of special education, multicultural education, health, physical education and art. Knowledge assessed by pedagogical items include cognitive and social development of students, lesson planning, classroom management, student evaluation, instructional strategies and student discipline.

Additional Validity Considerations

Knowledge Specific to Teachers

CONNECT items were designed to measure knowledge specific to teachers at a depth and breadth heretofore untapped. In effect, the examination was built to discriminate between prospective teachers trained specifically for that purpose from other candidates. Previous validity studies (Delandshere and Mason, 1989), using the contrasting groups approach (Berk, 1984), contributed evidence to support the construct validity of the examination. The results of these studies, which indicated a significant difference in the performance of teacher education and liberal arts students, were later confirmed by the performance of actual examinees. Administration of the first two forms of the examination resulted in differential pass rates of teacher education and liberal arts students (CT State Department of Education, 1991). A critical concern, therefore, was the impact of the differential performance of examinee subgroups on item parameters.

Curricular Alignment

Although teacher education program faculty and administrators were involved in assessing the content validity of the examination, some degree of misalignment was anticipated between what is taught and what would be measured. CONNECT sought to impose rigorous standards by measuring an understanding of major principles, philosophies, and documented effective practices in core subject areas of the elementary curriculum. In addition, since CONNECT is a licensure test, content decisions had to be based on the job-relatedness of items rather than the degree to which content had been incorporated into the teacher education curriculum.

By contrast to the intended test content, teacher education programs differed in program philosophy, focus and opportunity to learn. In effect, these programs reflected the ongoing inquiry and debate regarding the knowledge base for elementary education and how teaching principles were applied in the field. In an attempt to strike a balance, CONNECT was designed to be a catalyst for bringing teacher preparation curricula into line with critical skills that are valued on the job. It was hoped that as a result of CONNECT, opportunity to learn critical job-related skills would be incorporated into the curriculum.

Methodology

Instrument

Items selected for analysis were taken from two test forms of the CONNECT. These forms were administered prior to the date when the examination became a legal requirement for elementary education candidates. Therefore, in addition to a mismatch between curriculum and test content, incentive may have had some impact on test performance; therefore, the data analyzed may not be representative of later test performance.

Each form of the multiple-choice component of the examination contains the same proportion of items in each content area: pedagogy--15, reading and language arts--28, mathematics--22, science--10, social studies--10, multicultural education--5, special education--5, other content areas--5. Forty-one items represent common items; with the following distribution: pedagogy--6, reading and language arts--11, mathematics--9, science--4, social studies--4, multicultural education--2, special education--2, other content areas--3.

CONNECT is essentially an unspeeded examination with a total of 5 1/2 hours allotted for the completion of the entire examination; the constructed response portion of the examination is administered for approximately the first hour and a half, resulting in a minimum of four hours for the completion of the multiple-choice component of the examination. Each test form for this study was administered at five sites throughout Connecticut.

Respondents

A total of 826 responses were received from 681 adult candidates (aged 18 and over); 536 (78.7%) of the candidates responded to either form 1 or form 2 and 145 (21.3%) responded to both forms. The majority of candidates (approximately 90.3%) were enrolled in or graduates of one of Connecticut's 14 teacher education programs. Of the total number of candidates, approximately 1.0% were freshmen or sophomores, 44.8% were juniors, 30.0% were seniors, 13.3% had received their bachelors degrees, 2.6% were fifth year certification students,* and 8.3% had received greater than a bachelors degree.

Analyses

Analyses conducted comprised classical item statistics, including p-values and item discrimination (point biserials) for each multiple-choice form. Tests of fit were conducted using FACETS for the one-parameter model. Given the nature of the content of the examination, the possibility of meeting the assumption of equal discrimination, required for the one-parameter model, was unlikely. Correspondingly, guessing was anticipated in an examination of this kind. Due to these considerations, the data were also fit to the two and three-parameter models, using LOGIST, to determine the impact of these parameter estimates on difficulty.

Results and Analyses

Classical Item Statistics

Table 1 presents the mean performance, standard deviation, as well as the internal consistency for each form of the examination. As the data indicate, the difficulty of the forms remained relatively consistent, with performance declining slightly and variance increasing slightly on form 2. Internal consistency for each form of 0.83 and 0.84, indicating that the reliability coefficient may be affected by the multidimensionality of the test items.

Table 2 presents item difficulties as well as item discrimination for each form. P-values, the percent of candidates responding correctly to each item, indicate that the items ranged in difficulty from 0.24 to 0.97 on form 1 and 0.20 to 0.93 on form 2. Point biserials, item-total correlations, range from 0.02 to 0.44 and from 0.02 to 0.37 for forms 1 and 2 respectively.

P-values indicate a wide range of item difficulty, with the preponderance of items on both forms in the moderate range of difficulty. Since item statistics on criterion-referenced examinations may differ from those for norm-referenced examinations (Guiton and Delandshere, 1990; Millman and Greene, 1989, Millman, 1986), items with low p-values and point biserials were flagged and reviewed by content experts rather than automatically deleted. Review of the items resulted in the retention of all items for item response analyses.

Table 1: Descriptive Statistics for Form 1 and Form 2

Form	No. of Respondents	Mean	S.D.	KR-20
Form 1	423	65.21	10.60	0.83
Form 2				
Non-Retesters	258	62.58	11.06	0.84
All Examinees	403	61.09	10.21	0.81

Table 2: Classical Item Statistics By Form--
P-Values and Point-Biserials

Form	P-Values	Pt. Biserials
1		
9	00447	9
8	1122333588899	8
7	1222334555667789999	7
6	011222233334444666666888	6
5	00112223344555667788888899999	5
4	134569	4 44
3	35	3 000000001123444556688899
2	46	2 000001111122223333344445666677777888888889
1		1 0012224555556666777778888889999
0		0 2269
2		
9	0023	9
8	000022224467788	8
7	0245555778899	7
6	000112222244566688899	6
5	000012223444445555777788899	5
4	144555777788	4
3	2889	3 0000001122234557
2	023459	2 000000111122222233444445556666666777777888889999999
1		1 123344556666677777888899
0		0 246679

No negative point biserials were generated, supporting the conclusion that items discriminated appropriately, in favor of higher ability examinees. The concentration of point biserials in the 0.10 - 0.39 range suggests the possibility of heterogeneity; that is, test content may measure multiple underlying traits, violating the assumption of unidimensionality required for item response theory. Additionally, low point biserials may reflect the varying degrees of content coverage across teacher education programs alluded earlier; that is, the mismatch of teacher education program curricula and content knowledge validated for assessment. Although item statistics indicate multidimensionality, the previous identification of a dominant general factor (Guiton and Delandshere, 1990), indicates that the data adequately meet the assumption of unidimensionality (Hambleton and Swaminathan, 1985).

Item Response Theory Item Statistics

One parameter item analyses were conducted using FACETS (Linacre, 1988), a program derived from an extension of the Rasch model. FACETS was developed for the analysis of data with multiple facets. Accordingly, the program was designed such that categorical and dichotomous data might be placed on the same scale. A plot of parameter estimates produced by BICAL and FACETS resulted in a linear relationship between both estimations, further indicating equivalence between the values generated.

Fitting the One-Paramter Model A two-faceted model was specified (items and performance) for dichotomous data, using the Newton-Raphson procedure for joint maximum likelihood estimates. No items were eliminated due to the presence of maximum correct or incorrect scores. Accordingly, Rasch calibrations were conducted on all of the items.

Table 3 presents item difficulties, asymptotic standard errors and point biserials for selected items on forms 1 and 2 generated by FACETS. Item difficulties ranged from -2.75 (easy item) to 2.00 (hard item) for form 1, and from -2.07 (easy item) to 1.89 (hard item) for form 2. Mean difficulty estimates for both forms was 0.0, with S.D. = 0.83 for form 1 and S.D. = 0.85 for form 2. The mean asymptotic standard error for the difficulty estimates was recorded at 0.12 for form 1 (S.D. = 0.03) and 0.23 (SD = 0.02) for form 2. Estimates indicate increased model error in Form 2. However, model errors remained relatively low.

Mean point biserials were recorded at 0.19 for form 1 and 0.18 for form 2, with standard deviations of 0.08 and 0.07, respectively. The frequency of point biserials in the low range support previously documented evidence regarding possible heterogeneity of test content. By contrast to classical item statistics presented earlier, however, negative point biserials, -0.03 and -0.02 for items 8 (science--pedagogical knowledge) and 83 (science--content knowledge), respectively were generated by FACETS. The discrepancy in point biserials may be due to computational differences. However, these data provide preliminary evidence, later confirmed, that item 8 and 83 may not fit the model very well.

Table 3: Item Calibration from One Parameter Item Analysis

Form 1				Form 2			
Item	Item Dif.	Model Err.	Pt. Bis.	Item	Item Dif.	Std. Err.	Pt. Bis.
037	-2.75	0.27	0.13	079	-1.72	0.17	0.18
087	-2.17	0.21	0.09	031	-1.45	0.15	0.15
023	-1.56	0.17	0.13	053	-1.45	0.15	0.16
053	-1.36	0.15	0.17	081	-1.32	0.14	0.22
069	-0.93	0.13	0.36	073	-1.20	0.14	0.25
059	-0.90	0.13	0.35	027	-1.16	0.14	0.23
089	-0.88	0.13	0.18	025*	-1.05	0.13	0.20
025*	-0.81	0.13	0.11	083	-0.90	0.13	0.18
021*	-0.76	0.13	0.24	071	-0.87	0.13	0.27
045	-0.67	0.12	0.24	021*	-0.70	0.12	0.24
003	-0.61	0.12	0.22	035	-0.70	0.12	0.16
005	-0.55	0.12	0.34	001	-0.63	0.12	0.16
093	-0.53	0.12	0.41	019	-0.61	0.12	0.28
071	-0.48	0.12	0.14	039	-0.16	0.12	0.15
001	-0.47	0.12	0.24	095	-0.60	0.12	0.05
083	-0.43	0.12	-0.02	037	-0.34	0.11	0.11
035	-0.40	0.12	0.22	075*	-0.29	0.11	0.21
043*	-0.40	0.12	0.19	067	-0.24	0.11	0.08
007	-0.29	0.11	0.08	043*	-0.23	0.11	0.18
027	-0.05	0.11	0.27	061	-0.23	0.11	0.04
033*	-0.05	0.11	0.14	047*	-0.12	0.11	0.13
031	0.05	0.11	0.16	057*	-0.09	0.11	0.13
065*	0.06	0.11	0.26	029	-0.05	0.11	0.13
047*	0.08	0.11	0.18	089	0.04	0.11	0.26
013	0.17	0.10	0.30	045	0.05	0.11	0.20
015*	0.18	0.10	0.29	017	0.06	0.11	0.12
009	0.19	0.10	0.35	033*	0.11	0.10	0.21
017*	0.23	0.10	0.20	003	0.17	0.10	0.22
057*	0.29	0.10	0.19	049*	0.20	0.10	0.18
075*	0.36	0.10	0.31	065*	0.23	0.10	0.16
049*	0.38	0.10	0.20	085*	0.28	0.10	0.17
051	0.38	0.10	0.25	063*	0.33	0.10	0.27
097*	0.38	0.10	0.18	087	0.35	0.10	0.17
079	0.41	0.10	0.40	055	0.36	0.10	0.26
067	0.42	0.10	0.12	097*	0.38	0.10	0.24
091*	0.43	0.10	0.23	005	0.42	0.10	0.26
019	0.46	0.10	0.23	015*	0.46	0.10	0.25
063*	0.48	0.10	0.22	041*	0.54	0.10	0.31
029	0.55	0.10	0.22	011*	0.57	0.10	0.24
085*	0.59	0.10	0.14	093	0.65	0.10	0.32
011*	0.60	0.10	0.22	009*	0.68	0.10	0.21
073	0.67	0.10	0.32	069	0.68	0.10	0.21
039	0.68	0.10	0.25	091*	0.69	0.10	0.10
041*	0.70	0.10	0.17	023	0.78	0.10	0.15
009*	0.77	0.10	0.12	051	0.78	0.10	0.21
081	0.92	0.10	0.10	007	0.81	0.10	0.22
061	0.98	0.10	0.22	013	0.83	0.10	0.20
095	1.03	0.10	0.15	099	1.02	0.11	0.25
077*	1.15	0.10	0.24	077*	1.07	0.11	0.31
055	2.00	0.12	0.24	059	1.89	0.12	0.20

Infit statistics generated by FACETS is described as the information-weighted mean square statistic which has an expectation of 1 and a range of 0 to infinity. Outfit statistics is defined as the unweighted mean square which is more sensitive to outliers (Linacre, 1988). Infit mean squares for form 1 ranged from 0.9 to 1.1, with a standardized mean square range of -3 to 5. Outfit mean squares were equivalent to infit mean squares. Estimates generated indicate two items for which test statistics exceeded the absolute value of 3 (Linacre, 1988) and as possibly misfitting items--item 8 and item 79. Item 8, previously identified as having a negative point biserial generated by FACETS, addressed a general pedagogical principle regarding practice strategies and activities for newly-introduced topics. In item 79 (mathematics--pedagogical content knowledge), examinees were asked to identify a mathematics problem used to assess an understanding of the concept of sequence of operations.

Although item 8 was of moderate difficulty (p-value = .53), the item discriminated relatively poorly (classical: $r_{bis.}$ = 0.02, $P_{diff.}$ = 0.05). By contrast, while item 79 was also of moderate difficulty (p value = 0.58), the item discriminated extremely well (classical: $r_{bis.}$ = 0.44, $P_{diff.}$ = 0.51). These data indicate the items described were identified as misfitting because of their extreme discrimination ability--negligible ability to discriminate versus substantial ability to discriminate. Substantial discrepancy of point biserials from the mean and negative point biserials suggested that the items were very different from the majority of items on the examination; hence, the basis for their identification as misfitting items.

Although item 83 generated a test statistic between 3 and 2, the point biserial (-0.02) generated by FACETS indicates the item may also be misfitting due to extreme discrepancy of its point biserial from the mean. Classical item statistics, indicating patterns similar to the other misfitting items (p-value = 0.75, r_{bis} = 0.06, $P_{diff.}$ = 0.6). Item 83 taps knowledge of content (physical science--weather), covered by the elementary curriculum.

In accordance with form 1 estimates, infit mean squares for selected items on form 2 ranged from 0.9 to 1.1, with a standardized mean square range of -2 to 3. Again, outfit mean squares were equivalent to infit statistics. Based on the guidelines of std mnsq greater than or equal to 3, items 28 and 90 were identified as misfitting items. Closer examination of classical item statistics again revealed characteristics similar to those for misfitting items on form 1; the items were of moderate difficulty that don't discriminate very well (item 28--p-value = 0.50, $r_{bis.}$ = 0.04, $P_{diff.}$ = 0.07; item 90--p-value = 0.54, rbis. =0.06, Pdiff. = 0.06). Both items items address knowledge of reading and language arts pedagogical principles; specifically, process writing and word identification strategies, respectively.

In general, the content of the misfitting items varies across subject area and knowledge type. Judgmental evaluations of the items revealed no fundamental item content concerns. However, the content of the items suggested that correct responses could depend on opportunity to learn current understandings about specific instructional strategies, content as well as the application of particular strategies in certain content situations. These explanations were further supported by parameter estimates for the common items.

Item 28 is one of the 41 items that are common to both forms 1 and 2. These items serve as a base item pool for equating purposes. Performance on this item, together with a number of common items (denoted by asterisks in Table 3), generated small discrepancies in item statistics across forms, with items shifting in both directions. On form 1 performance on item 28 resulted in a difficulty (b) estimate of 0.30, while on form 2 the same item resulted in b = 0.56, an increase of 0.26. Mean difficulty estimates for the common items on forms 1 and 2 resulted in a difference of 0.08, with a maximum decline in difficulty of 0.65 logits and a maximum increase of 0.29 logits.

Although the shifting of difficulty estimates are extremely small, they are incongruous, given the item response assumption of invariance of estimates across samples. These discrepancies were confirmed by another program (TRION) designed for the one parameter model. Notably, however, the differences between mean estimates for non-retest examinees reveal a small reduction in the size of the discrepancy from 0.08 to approximately 0.023.

The decline in the discrepancy of parameter estimates may support the contention that discrepancies are content based; that is, it may support the argument that test content and teacher education curriculum may be misaligned, since estimates based on the entire pool reflect to a greater extent those who've had less opportunity to learn. Linn (1990) suggests that the lack of invariance of common items across forms may be due to differential opportunity to learn as well as recency of instruction. He notes that studies where substantial differential effects have been found focused on students at higher grade levels with great differences in the kinds of instruction experienced. Regarding the implication of these kinds of results, Linn (1990) concludes, "If items that are found to be most sensitive to instruction are eliminated so that IRT assumptions are better satisfied, there is a real danger than IRT will do more to decrease than to increase validity...Such an outcome could have very negative consequences for the educational uses of achievement tests" (p. 136). Judgmental evaluations of the common items identified as lacking invariance revealed no fundamental deficiencies with regard to item content.

Fitting the Two Parameter Model. Default options were used to fit the data to the two-parameter model. Form 1 generated a mean difficulty estimate of b = -0.87, S.D. = 0.95, while form 2 resulted in a mean difficulty estimate of b = -0.89, S.D. = 1.09.

Form 1 yielded an extremely wide range of discrimination estimates, from 0.01 to 1.00. Two misfitting items (items 8 and 83) were estimated to have a discrimination estimate of 0.01. These items were also identified as potentially failing to fit the one-parameter model. Item 79, the other item identified as inadequately fitting the one-parameter model, fit the 2-parameter model very well, possibly due to the high discriminating ability of this item (a = 0.78). Additional misfitting items for the two-parameter model were, in general, those items failing to discriminate, specifically items 18 (a = 0.078), 24 (a = 0.02) and 50 (a = 0.09). It is notable that items failing to fit the two-parameter model also had extreme difficulty estimates; for example, items 24 and 83 registered difficulty estimates of -73.21 and -65.85, respectively. It is not clear at this time whether the identification of these items as misfitting may be due to the characteristics of the items or to the inability of LOGIST to compute item parameter estimates of certain items.

The range of item discrimination estimates for form 2 were more moderate than those for form 1, with a = 0.01 to 0.63. Item 61 was identified as having a discrimination estimate of 0.01, with b = -43.56. Classical item statistics indicated a relatively easy item (p-value = 0.68), with poor ability to discriminate. However, it is somewhat puzzling that item 61 appeared to fit the one-parameter model. Items 28 and 90 identified as having questionable fit for the one-parameter model appeared to fit the two-parameter model fairly well, with moderate to easy item difficulty as estimated by the two-parameter model (b =0.06 and b = -2.31, respectively). In accordance with form 1, items 2, 46 and 76 yielded extreme positive or negative difficulty estimates, with poor discrimination (a = 0.05, 0.06 and 0.07, respectively).

Review of items 46 and 76 revealed two special education items. Both items required an understanding of the legal guidelines surrounding the diagnosis of learning or physical disabilities from academic performance. Both items required the review of reference materials. However, no other misfitting item required the review of reference materials.

Item parameter estimates were examined for invariance of common items. Discrepancy in parameter estimates across forms were also noted here, with a discrepancy of 0.25 between the means of the difficulty estimates for those items (form 1: mean b = -1.03, form 2: mean b = -0.78). Shifts in item difficulty occurred in both directions as they did using the one-parameter model.

Fitting the Three-Parameter Model Convergence did not occur for the three-parameter model. Although it is not clear why the data did not fit the model, some proposals may be extended. The parameter c, also called the pseudo-chance level parameter in the three-parameter model, "accounts for item response data from low-ability examinees where ... guessing is a factor in test performance " (Hambleton and Swaminathan, 1985). Explanations for the failure of the item response data to converge when fitted to this model may include inadequate response data reflecting low ability, somewhat disfunctional distractors or inadequate sample size. All of these are feasible given the limited sample size used as well as other limitations previously described concerning the timing of the administrations.

Discussion

At the beginning of this paper, the unconventional characteristics of CONNECT items were noted, particularly the multidimensionality of the test items and the multiple types of knowledge measured. These, together with the variability of Connecticut teacher education program curricula and the relatively modest sample sizes particularly for the two-parameter model, support cautious interpretation of the results presented. Additionally, it is important to note that due to the timing of the administrations, the results presented may be atypical.

Classical item statistics indicate test items that are relatively heterogeneous, with the highest frequency of point biserials in the 0.10 to 0.30 range. These statistics confirm previous findings of more than one underlying construct. However, subsequent successful attempts to fit the data to one and two parameter models support the contention that a dominant underlying factor is sufficient to meet the assumption of unidimensionality required for item response analyses.

Item statistics for the one and two-parameter models suggest that forms 1 and 2 fit these models fairly well. In addition, results indicate that those items failing to fit either model adhered to a specific pattern of discrimination and difficulty. In sum, while items failing to fit the one-parameter model tended to have extremely high or low discrimination estimates, misfitting items for the two-parameter model tended to have poor discrimination ability and extreme difficulty estimates.

While the two-parameter model provides additional information necessary to evaluate items built to tap teacher-specific knowledge, extreme difficulty estimates are a concern. Further investigation will be necessary to determine the source of extreme estimates as well as the model providing the better fit of response data. Specifically, the relationship between difficulty estimated by the one and two-parameter models may be examined to determine if a nonlinear relationship exists; such evidence may suggest that extreme estimates computed by LOGIST may reflect differential impact. Regardless of the outcome, efforts should continue to select items which discriminate around the cut score.

Although much more exploration is needed, evidently the shifting of the logits may very well be explained by a lack of curricular alignment with test content (Linn, 1990). The reduced discrepancy noted when item parameter estimates are obtained from first-time examinees, suggest that additional information may be derived from plots of item parameter estimates of other subgroups, including subgroups containing respondents completing student teaching versus those who haven't.

While misfitting items exhibited a pattern of psychometric qualities, no content-based pattern was detected. Items failing to fit the one- and two-parameter model did not consistently fall in specific categories of knowledge measured, particular content areas, nor were they characteristic of particular item formats. These findings seem to support the notion that these misfitting items may reflect differences in the focus of teacher education curricula and differences in opportunity to learn. These issues of validity will continue to be explored. Specifically, studies addressing the relationship between test performance and opportunity to learn will be conducted. One method under consideration is a structural equation model specifying the relationship among various various factors, including opportunity to learn and aptitude, and performance.

APPENDIX A

CONTENT KNOWLEDGE:	The knowledge, understanding and skills to be learned by schoolchildren. Content knowledge rests on two foundations: the accumulated literature in the content areas and the philosophical scholarship on the nature of knowledge in those fields.
PEDAGOGICAL KNOWLEDGE	The knowledge of teaching, including generic principles of organization, management and other principles of teaching that apply to all content areas.
PEDAGOGICAL-CONTENT KNOWLEDGE	The blending of content and pedagogy into an understanding of how particular topics, problems or issues are organized, represented, and adapted to the diverse need of learners, and presented for instruction.

Adapted from Shulman, L.S. (1987) Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57, 1-8.

REFERENCES

Berk, R. (1984). A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press.

Delandshere, G. & Mason, T.C. (1989). The effect of opportunity to learn on test item statistics. Paper presented at the Annual meeting of the National Council on Measurement in Education, San Francisco.

Guiton, G. W. & Delandshere, G.(1990). Content-specific pedagogical test items: Measurement issues. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Linacre, J.M. (1988). FACETS: Many faceted Rasch analysis. Chicago: Mesa Press.

Linn, R.L. (1990). Has item response theory increased the validity of achievement test scores? Applied Measurement In Education, 3 (2), 115-141.

Millman, J. (1986). Reliability. In Jaeger, R.M. & Busch, J.C. (EDS.) Evaluation of the Georgia Teacher Certification Testing Program. Greensboro: Center for Educational Research and Evaluation.

Shulman, L.S. (1987) Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57, 1-8.

Hambleton R. K. & Swaminathan, H. (1985). Item Response Theory: Applications and Applications. Boston: Kluwer Nijhoff Publishing.